UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

MÁSTER EN BIOLOGÍA COMPUTACIONAL

DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL (ETSIINF)

DEPARTAMENTO ONCOLOGIA MOLECULAR (HCSC)

Integrative Proteomic Analysis of Breast Adipose Tissue-Derived Extracellular Vesicles: Insights into Differentially Expressed Proteins and Their Role in Cancer Pathway Activation

TRABAJO FIN DE MÁSTER

Autor: Laura V. Piñeres Santos

Tutor Externo: Alberto Benito Martín

Tutor Académico: Bojan Mihaljevic

July, 2025

ACKNOWLEDGMENT

To the members of Oncology laboratory at Hospital clinic San Carlos (HCSC) for their great collaboration and supportive environment.

To my family and friends who have always supported me, even more so from a distance, especially my nephew Tomas for his patience and love while waiting for me.

Thanks to all of you.

INDEX

FIGURES INDEX

TABLE INDEX

ABBREVIATION LIST

AT	-	Adipose tissue
ATME	-	Adipose tissue microenvironment
BC	-	Breast cancer
BCC	-	Breast cancer cells
BMI	-	Body mass index
DDA	-	Data dependent analysis
DEA	-	Differential expression analysis
DEPs	-	Differential expression proteins
DT	-	Desition tree
ER	-	Estrogen
EVs	-	Extracellular vesicles
FC	-	Fold Change
FDR	-	False discovery rate
GO	-	Gene ontology
IPA	-	Ingenuity pathway analysis
KNN	-	K nearest neighbor
LR	-	Logistic regression
MAE	-	mean absolute error
ML	-	Machine learning
MS	-	Mass spectrometry
O	-	Obese
OW	-	Overweight
PCA	-	Principal component analysis
PPI	-	Protein-protein interaction
RF	-	Random forest
RMSE	-	Root mean squared error
ROC	-	receiver operating characteristic
SI	-	Silhouette score
TME	-	Tumor microenvironment
UMAP	-	Uniform manifold approximation and projection

ABSTRACT

Obesity is a known risk factor for breast cancer, yet the molecular mechanisms linking adipose tissue dysfunction to tumor progression remain incompletely understood. Extracellular vesicles derived from breast adipose tissue are a key mediator of cell-to-cell communication within the tumor microenvironment, capable of influencing cancer cell metabolism and proliferation. In this study, we applied an integrative proteomic and computational pipeline to explore the differential expression and predictive significance of EV-associated proteins derived from breast adipose tissue in overweight and obese women.

An open-source proteomic dataset of EVs isolated from breast adipose tissue was reanalyzed and filtered for high-confidence peptides with experimental metadata. Preprocessing included $\log_2$ transformation, imputation, and feature scaling. Dimensionality reduction was performed using a Principal Component Analysis (PCA) + UMAP hybrid strategy, followed by unsupervised K-means clustering, which revealed two molecularly distinct sample clusters not associated with any previously noted metric. Differential expression analysis identified 665 proteins, including SERPINA1, CAV1, MMP9, and AKR1C2, with significant enrichment in extracellular matrix remodeling, hormone metabolism, and immune modulation. A supervised machine learning approach using Decision Tree regression predicted breast cancer cells (BCC) proliferation from EV proteomic profiles (R^2 = 0.87, AUC = 0.98 for binary classification). Top-ranked predictive proteins, including PARP1, FCN2, and HLA-G, were further explored for functional relevance, although not all overlapped with canonical pathways explored by a comprehensive systems biology software (IPA-Ingenuity Pathway Analysis) which identified Estrogen-Dependent Breast Cancer Signaling and Tumor Microenvironment as canonical pathways, while CTNNB1-centered node network analysis revealed EV-associated proteins with known roles in metabolic rewiring and tumor progression. Further validation is needed to enhance the robustness of the machine learning model for early breast cancer risk stratification and molecular subtyping

Keywords: *Obesity, cancer-related, Pathways, EVs, Machine Learning.*

1. INTRODUCTION

1.1 Obesity and cancer

Obesity is a growing global health concern, affecting approximately one in eight individuals worldwide according to the World Health Organization (WHO, 2022). Classification is primarily based on body mass index (BMI), where individuals with a BMI between 25–29.9 kg/m² are considered overweight (OW), and those with a BMI of ≥ 30 kg/m² are classified as obese (O). A large-scale epidemiological study conducted between 1980 and 2015, encompassing data from multiple countries representing 78% of the global population, reported a 50% increase in obesity prevalence among women during this period (Chooi et al., 2019). The etiology of obesity is multifactorial, involving socio-economic status, sex, age, lifestyle, genetic predisposition, and epigenetic factors (Ferreira et al., 2021). Recent attention has also been drawn to the role of ultra-processed food consumption as a significant contributor to the obesity epidemic (Marti et al., 2021). Obesity is associated with various metabolic and physiological disturbances, including chronic low-grade inflammation, insulin resistance, and altered hormone signaling. These changes contribute to the pathogenesis of multiple comorbid conditions such as type 2 diabetes mellitus (T2DM), cardiovascular disease, and non-alcoholic fatty liver disease (NAFLD) .(Beckman & Creager, 2016; Mokdad et al., n.d.; Yaturu & Jain, 2007, 2007)

In recent years, an increasingly complex, well studied, but not fully understood relationship between obesity and cancer has been described. Obesity is now recognized as a major risk factor for at least 13 types of cancer, including breast, colorectal, endometrial and pancreatic cancers (Gathirua-Mwangi et al., 2017; Kyrgiou et al., 2017; Macinnis et al., 2004; Nevadunsky et al., 2014; Nwafor et al., 2025). According to epidemiological estimates, approximately 40% of all cancers diagnosed annually in the United States are associated with OW or O (Nwafor et al., 2025).

In breast cancer (BC), there is a limited literature exploring the association between disease progression and obesity (James et al., 2015). However, BC is the most prevalent malignancy among women worldwide, exhibiting a rising incidence, particularly among younger populations. This trend underscores the critical need for a comprehensive understanding of its diverse subtypes, each characterized by distinct molecular and histopathological features that influence prognosis and therapeutic strategies [(Hong & Xu, 2022)]. Molecular classification has delineated breast cancer into several subtypes: Luminal A, Luminal B, HER2-enriched, and triple-negative (basal-like). Luminal A tumors are typically estrogen receptor-positive (ER+), progesterone receptor-positive (PR+), HER2-negative. These characteristics confer a favorable prognosis and responsiveness to hormonal therapies. In contrast, Luminal B tumors are also ER-positive but may be HER2-positive and have a comparatively poorer prognosis. In the other hand, HER2-enriched tumors are associated with aggressive disease prognosis but with treatment available, contrary to triple negative (TN) characterized by the absence of hormone receptors, and HER2 protein, making it more aggressive and less responsive to currently available therapies (Park et al., 2012).

Studies linking obesity and breast cancer attempt to find an indicative metric that may

influence tumor biology and patient outcomes. While BMI is a commonly used metric for obesity, it may not fully capture the complexity of obesity-related cancer risks. However combining metrics like waist circumference, fat mass has been associated with risk cancer 15 years after menopause (Macinnis et al., 2004). In-depth molecular analyses provide more reliable insights into this connection, presenting a complex multilayered influence, in which multiple events participate (Avgerinos et al., 2019; Bowers et al., 2015). Some of the key pathways involved are hyperinsulinemia/IR and abnormalities of the insulin-like growth factor-I (IGF-I) system and signaling, sex hormones biosynthesis and pathway, subclinical chronic low-grade inflammation and oxidative stress, alterations in adipocytokine pathophysiology, factors deriving from ectopic fat deposition, microenvironment and cellular perturbations.

1.1.1. Adipose tissue as a key biological regulator in breast cancer

Adipose tissue (AT) traditionally recognized for its role in energy storage through adipocyte hypertrophy and hyperplasia, is also implicated in breast development. Its mass increases in obese individuals ; (Quail & Dannenberg, 2019; Rosen & Spiegelman, 2014) and is now considered a dynamic endocrine organ. AT can influence tumor behavior at molecular and cellular levels through its quality and secretory profile, contributing to a specialized adipose tissue microenvironment (ATME). This environment is shaped by the release of various bioactive molecules, including hormones, cytokines, chemokines (e.g., CCL2, CCL5), adipokines (such as leptin and adiponectin), and free fatty acids (FFAs), which can drive metabolic reprogramming in cancer cells.

Normal adipocytes are driven into cancer-associated adipocytes by tumor cells and at the same time hijack surrounding metabolic pathways ; (C. Wu et al., 2023; Q. Wu et al., 2019). In order to supply the extreme needs of energy of dividing cells, process like adipogenesis differs into lean and O-OW individuals by alteration in the metabolism of proteins, carbohydrates and lipids (Cairns et al., 2011). Some already identified markers involved in these processes are PPARγ named as master regulator of fat cell proliferation, and several transcription factors including C/EBPα, C/EBPβ, and C/EBPδ (Rosen & Spiegelman, 2014). Estrogen (ER) biosynthesis can be generated from the adipose tissue by the production of aromatase, a key enzyme for ER increasing the circulation of the hormone (Gérard & Brown, 2018). In BC it promotes the growth and survival of hormone receptor-positive tumor cells by activating estrogen receptors that regulate gene expression involved in cell proliferation. Elevated estrogen levels or sustained signaling can contribute to tumor initiation, progression, and resistance to therapy (Yoon et al., 2022). Finally, understanding the interactions involving AT biology opens avenues for therapeutic targeting, such as inhibiting adipogenesis, modulating the adipose secretome, and interfering with signals from the tumor microenvironment (TME) (Kothari et al., 2020). Among these signals, EVs derived from AT have emerged as key mediators of crosstalk, carrying bioactive molecules—including proteins, lipids, and nucleic acids that influence cancer cell behavior. Studies have shown that EVs from obese adipose tissue (O-EVs) can promote BCC proliferation and migration (C. Zhou et al., 2023).

1.1.1.1. O-EVs altering metabolism

EVs are delimited by lipid bilayer playing crucial roles in intercellular communication. Their cargo reflects the physiological or pathological state of their cell of origin and its capacity to trigger different pathways in the recipient cell (Lilite Sadovska et al., 2015). The International Society for Extracellular Vesicles (ISEV) has provided guidelines for the standardization of EV nomenclature and characterization, which are classified based on their size and biogenesis into: Exosomes: 30–150 nm vesicles originating from the endosomal pathway and Microvesicles: 100–1000 nm vesicles formed by direct budding from the plasma membrane (Welsh et al., 2024). Exosomal contents modulate cell biology by trafficking material like mRNA, ncRNA and transcription factors.
EVs produced by certain cell types express cell-type specific markers, and the amount of EVs circulating in plasma detected in oncological patients increases in relation with healthy patients, similarly with lean and O-OW individuals, (Logozzi et al., 2009; Matilainen et al., 2024). It remains interesting to know whether the cargo of EVs also shows a similarity that helps to predict the prognosis of obese individuals. What has been described so far is that elevated levels of EVs are associated with therapy failure and disease progression in BC patients undergoing neoadjuvant chemotherapy (König et al., 2018). Liu S, and Benito-Martin, A. has demonstrated that O-EVs induce metabolic alteration in ER+ MCF7 and T47D BC cells lines by increasing the expression of genes involved in oxidative phosphorylation. (Liu et al., 2023) They also reported several miR (miR-373, miR-101, miR-372, miR-155-5p, miR-10a-3p, and miR-30a-3p) involved in Akt/mTOR/P70S6K signaling pathway . Other glycoproteins like CD147 and GPC1 is also enriched in EVs (Eichelser et al., 2014; Liu et al., 2023; Melo et al., 2014; Menck et al., 2015). Furthermore, O-EVs can modulate the immune landscape of the TME, potentially leading to immune evasion and therapy resistance reported by the expression of vascular endothelial growth factor (VEGF) that is sensible to Bevacizumab (Feng et al., 2017; Kumar et al., 2024). O-EVs are positioned as a potential BC biomarker source using novel omics techniques for their cargo identification.

1.2. Computational approaches in proteomic analysis

Proteomics is the large-scale study of proteins, including their structures, functions, modifications, interactions, and expression levels within a biological system. It is part of the novel omics techniques allowing identification and quantification of protein from samples using different techniques. Traditionally, mass spectrometry (MS), particularly tandem MS/MS, has been the predominant technique for proteomic analyses (Medzihradszky & Chalkley, 2015). However, novel approaches are developed constantly with different identification principles (Callahan et al., 2020). The data acquisition employs two strategies: data-dependent acquisition (DDA) that selects as many peptides as possible generating missing values, and data-independent acquisition (DIA), a method that produce a complex MS/MS spectrum. To assess relative protein abundance, quantification can be performed using either label-free methods (LFQ), which

compare raw data across proteomic conditions based on signal intensity, or label-based approaches, which incorporate stable isotope labels to allow for relative or absolute quantification. (Matthiesen & Bunkenborg, 2013). The following steps for database searching and analysis used to be performed by software-based tools such as MaxQuant, Proteome Discoverer and X!tandem (Schmidt et al., 2014).

Raw proteomic data is inherently noisy, incomplete, and high-dimensional, making it impossible to extract meaningful biological insights without proper preprocessing and quality control measures. Some steps to improve the quality implies statistical analysis for pattern recognition and classification to add biological meaning. A normalization of intensities or ratios with imputation are performed to handle outlier, different scales and missing values (Holger Husi & SC Nat, 2019) to generate a complete matrix that preserve the information and avoid bias in the follow analysis. To reduce high-dimensionality, different methods can be use such as, principal component PCA by orthogonal classification, Uniform Manifold Approximation and Projection (UMAP), and t-distributed Stochastic Neighbor Embedding (t-SNE), some researchers suggest a mixture model to achieve a proper reduction without creating an artifact by tuning the hyperparameters , (Becht et al., 2019; Hilario & Kalousis, 2008). These approaches lay the groundwork for the next clustering, used to find expression patterns of groups of proteins and visualize the results, using the most popular distance measurement based on the Pearson Correlation (PE) coefficient (1 - r) or Euclidean metric (EU) methods. Both are affected more strongly by very abundant protein than the low-abundance use in hierarchical clustering analysis (Meunier et al., 2007). Other methods are extensively used like K-means and novel algorithms that successfully denoise proteomic spectra called Bayesian-Fourier model based (Bensmail et al., 2005). The identified cluster must be statistically evaluated with strategies using Silhouette coefficient (SI) and statistical models are applied for differentially expressed proteins by univariate methods including t-test and analysis of variance (ANOVA) (Lualdi & Fasano, 2019).

Proteomic analysis facilitates the characterization of complex biological systems, particularly in cancer research, by enabling the identification of protein-level alterations associated with obesity. The analytical power of proteomic studies can be enhanced through optimal parameter selection and comparative methodological approaches, which increase the likelihood of obtaining statistically significant results while minimizing experimental bias, thereby allowing researchers to select the most appropriate analytical strategy for their specific research objectives and improving predictive and preventive capabilities in translational medicine applications.

1.3. Predictive modeling in risk cancer

Initial proteomic analysis combined with machine learning (ML) algorithms enables robust data integration and enhanced extraction of the hidden message within complex datasets, facilitating the inference of biological properties from training datasets to predict clinical conditions based on existing data. While conventional statistical approaches were previously employed to analyze factors influencing BC, these traditional methods exhibit limited adaptability for discovering novel variables and generating integrative data

visualizations (Ganggayah et al., 2019). ML improves the ability to analyze and generate accurate outcomes used in translational medicine, and it is done by identifying proper correlation like obesity-associated parameters that can influence or be responsible for BC development.

A predictive model development follows a workflow starting by model training, employing machine learning algorithms, model validation and testing, and finally, interpretation and deployment. ML algorithms extensively used in cancer proteomic based models are random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM) and artificial neural network (ANN) , (Al Mudawi & Alazeb, 2022; Kourou et al., 2015) that need to split the initial dataset into subsets to be employed through proportional partitions. Subsequently, model validation is essential, employing cross-validation techniques like K-folds to asses ability and prevent overfitting (Kourou et al., 2015). Performance evaluation utilizes multiple metrics including accuracy, precision, recall, F1-score, and particularly the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), which summarize the trade-off between sensitivity and specificity across thresholds (LG & AT, 2013). Accuracy evaluates all the correct predictions (true positive and true negative) over all the predictions (all true and false positive and negatives).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

ML allows the integration of diverse features such as proteomic expression profiles, obesity-related parameters, and clinical metadata into models that can predict disease risk, subtype classification, or treatment response with high reliability. Recent studies have proposed models based on the predictive capability for BC prognosis by integrating BRCA1/2 genetic mutations with lifestyle-related risk factors, thereby exhibiting robust discriminatory power to differentiate between individuals at high and low risk development; the models are summarized by Conte L., et. al., in 2024 with an average AUC of 81% (Conte et al., 2024), incorporating obesity related factors to the model can potentially design a smart BC risk prediction for all the associations extensively described. These models present the potential for an early identification stage and preventive strategies.

1.4. Biological interpretation and functional analysis

Beyond statistical modeling, interpreting the biological significance of differentially expressed proteins is crucial, providing a comprehensive understanding of the contribution of various biological processes related to obesity and possible BC. This analysis is possible thanks to research community collaboration for annotating proteins with their biological function, pathways and interactions to provides these insights tools such as Gene Ontology (GO) enrichment that categorize proteins based on their associated (i) biological processes, (ii) molecular functions, and (iii) cellular components (Ashburner et al., 2000).

Proteins identified through differential expression analysis can undergo pathway analysis

using various bioinformatics platforms designed for multi-omics applications. Widely used approached include KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis , , (Chen et al., 2015; Neagu et al., 2021; T. Zhou et al., 2017), and other tools known as GOMiner (Zeeberg et al., 2005) and PANTHER, which provide computational resource automatization for omics data by a classification system that enhanced the robustness by allowing comparison with existing datasets. However, one of the most used and complete tools for interpretation is Ingenuity Pathway Analysis QIAGEN software (IPA) GO and biological interactions to predict upstream regulators and downstream effects (Krämer et al., 2014).
These analyses enable the contextualization of proteomic alterations within known biological frameworks, aiding in the identification of potential biomarkers and therapeutic targets. There are several conducted studies using different ML algorithms, however few of them consider protein source from O-EVs to mapped the increasing risk relationship with BC.

2. MAIN OBJECTIVE

- Understand how EVs derived breast adipose tissue from overweight and obese women may reflect or influence cancer-related pathways.

2.1. SPECIFIC OBJECTIVES

- Identify differentially expressed proteins in EVs derived from breast adipose tissue.
- Develop and select different machine learning models to classify patterns within the proteomic data that differentiate groups of patients based on their EV-derived protein profile and cell proliferation index.
- Determine the metabolic pathways of the enriched proteins to explore their biological functions and implications in breast cancer by Ingenuity Pathway Analysis.

3. MATERIAL AND METHODS

All computational analyses were performed using a MSI Creator A16 workstation equipped with RTX 4060/16 GPU. The analytical pipeline was implemented in JupyterLab using Python 3.10 with scikit-learn packages. All source code and documentation are publicly available at https://github.com/lvpineres/TFM_2025/

3.1. Dataset acquisition and description

Proteomic data were obtained from a previously published study (Liu et al., 2023) investigating breast adipose tissue-derived extracellular vesicles from obese women. The

original study collected breast adipose tissue samples from 48 overweight and obese women undergoing elective breast reduction surgery at Weill Cornell Medicine, New York. EVs were isolated from these tissue samples following the protocol described in the original publication. The authors conducted proteomic profiling using MS/MS for all the samples alongside, proliferation assays on MCF7 breast cancer cell lines only to 81% of the total collection.

The corresponding proteomic data remained largely unexplored. The dataset used for this analysis is publicly available via the PRIDE Archive repository under the accession number PXD045471 (https://www.ebi.ac.uk/pride/archive/projects/PXD045471). The data was originally processed using MaxQuant software forDDA proteomics and exported as proteinGroup.txt, however the file used was in in CSV format including the key varia bles [Intensity values per sample, Protein IDs and names, Gene names, FASTA headers, Number of peptides identified, Sequence coverage (%), Molecular weight (kDa), Q-values]. Additionally, clinical metadata, including age, Body Mass Index (BMI) and cellular proliferation index were available for each donor sample.

3.2. Data preprocessing and quality control

To optimize data preprocessing for downstream analysis, a systematic comparative evaluation was conducted using SI validation. Two datasets were analyzed: the complete dataset and a filtered subset excluding low-quality peptides with ≤ 20% sequence coverage. Both datasets underwent $\log_2$ transformation of intensity values, followed by missing value imputation using three methods: leave-tail imputation, simple imputation, and iterative imputation. Distribution analysis guided the selection of the optimal imputation strategy, which was subsequently applied with different normalization approaches (StandardScaler, MinMaxScaler, and RobustScaler) to standardize feature ranges. After validation of data distributions, dimensionality reduction techniques were systematically compared, including PCA and UMAP with both default and optimized parameters (varying numbers of neighbors and minimum distances) until achieving clear separation into two distinct sample groups. The dataset was transposed using df.T function (features as columns and samples as rows) to facilitate unsupervised clustering using K-means and Agglomerative approach. Final cluster configurations were chosen based on visual distribution inspection of separation and SI.

3.3. Differential expression Analysis (DEA)

For DEA, a pre-built package was applied to the dataset. To select the most suitable package according to the literature review, 7 initial models were compared by applying direct filtering methods based on their availability, community support and peer-reviewed documentation for the preliminary selection of 3 models, new filters were applied based on operational parameters and ease of installation and configuration. The pipeline selection is shown in **Figure 1.** A final open-source package was carefully selected called AlphaPepStats (Krismer, E. *et. al*., 2023) that allows the MaxQuant pre-processed file

with the metadata information and its automatization.

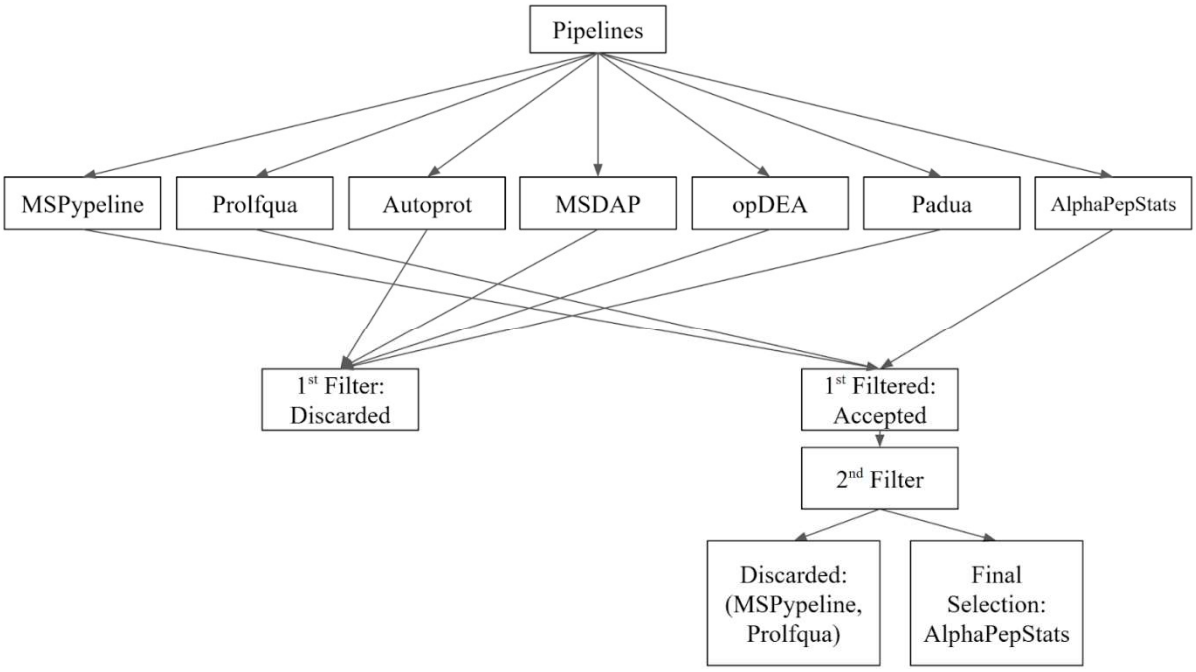


Figure 1. Hierarchical and selected pipeline in different filtering steps.

Its image was pulled using 'docker run -p 8501:8501 elenakrismer/alphapepstats_streamlit' to identify proteins that were differentially expressed between clusters and ensure compatibility AlphaPepStats pipeline, the proteinGroup.txt file in MaxQuant format was used as the input dataset. To maintain integrity the LFQ intensity columns were isolated and subjected to quality control and preprocessing steps already described in section 4.2, the column was reintegrate into the original dataset using the df.loc[]. Statistical analysis was conducted such as ANOVA across sample groups applying a significance threshold of $p\ value < 0.05$ and $\log_2$ fold change (FC) cutoff of ± 1.Similarly, a custom de novo pipeline was developed to perform DEA on the labeled, preprocessed proteomic dataset. A two-sided independent Student's t-test was applied by calculating Mean difference (Mean Diff) in LFQ intensity between groups, following the criteria $P\ value > 0.05$,FC calculated as the $\log_2 FC \geq |0.25 - 2|$ and False discovery rate (FDR) applied using Benjamini-Hochberg [method='fdr_bh'] procedure to control for multiple testing error using libraries like [itertools, scipy.stats and statsmodels.stats.multitest]. Differential expressed proteins (DEPs) over these thresholds are retained in an annotated list and visualize in a volcano plot generate using matplotlib.

3.3.1. Protein-protein interaction network analysis

To understand the functional relationships between differentially expressed proteins, a protein–protein interaction (PPI) network analysis was constructed using the STRING database https://string-db.org/ from the retained list that include the statistically differentiate proteins across clusters. Separate interaction networks were generated for

upregulated and downregulated proteins setting *Homo sapiens* as organism. Minimum required interaction score was set to high confidence (score > 0.7) to exclude weak or speculative associations and enhance the robustness of the resulting network. This DEPs were used to constructed the final interaction network. A summarization of the steps as shown in **Figure 2.**

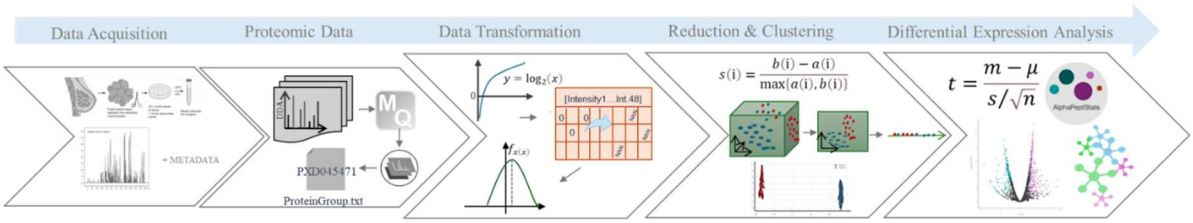


Figure 2. Workflow of data preparation and analysis. The first 2 steps shown in grayscale were performed by the authors in the original study and color steps summarize the strategies for data curation.

3.4. Machine Learning Model Development and validation

Four machine learning algorithms were selected for evaluation based on their wide use in biomedical data analysis and their ability to handle complex, high-dimensional proteomic data. The preprocessed data frame obtained from previous steps is used for determining the proliferation index of breast cancer cells based on the expression profiles of proteins derived from extracellular vesicles in adipose tissue included in the metadata.

For model comparison, comprehensive parameter optimization was performed across all algorithms, with the best-performing configuration selected for each method. The Random Forest model comprised n_estimators [20, 50, 100, 150, 200], max_depth [3, 5, 10, None], min_samples_split [2, 3, 5, 10, 15, 20]. For the Decision Tree (DT) model, cost-complexity pruning was implemented based on cross-validation results by comparing max_depth [3, 5, 10, None], min_samples_spli [2, 5]. Logistic Regression (LR), traditionally employed for classification tasks, was adapted for regression analysis using a logit link function with L2 regularization compared in different alpha [0.01,0.1, 1, 10, 100] to predict continuous proliferation indices. Additionally, we also compared K-nearest neighbor n_neighbors [3, 5, 7] weights [Uniform, distance] metric [euclidean, manhattan].

The evaluation of each model was performed by 5-fold cross-validation to check model stability and the performance metrics were Root mean squared error (RMSE) that measures the average squared difference between predicted and actual proliferation indices, R-squared (R^2) to indicate the proportion of variance and MAE to show mean absolute error. The models were compared based on the above metrics, with the best-performing model exhibiting the lowest RMSE, MAE and the highest R^2, a summarization is found in **Figure 3.**

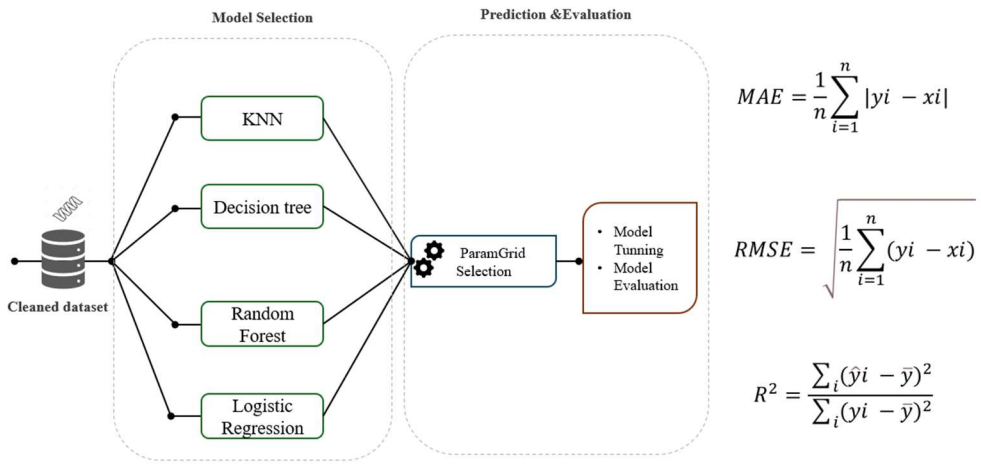


Figure 3. Model comparison and evaluation. Input dataset with all preprocessed steps, the selected model includes the lower MAE and RMSE values and higher R^2, *yi* Represent actual values and *xi* predicted values with the n = number of total observations, $\bar{y}$ represent the mean of *y*i values, and $\hat{y}i$ are predicted values.

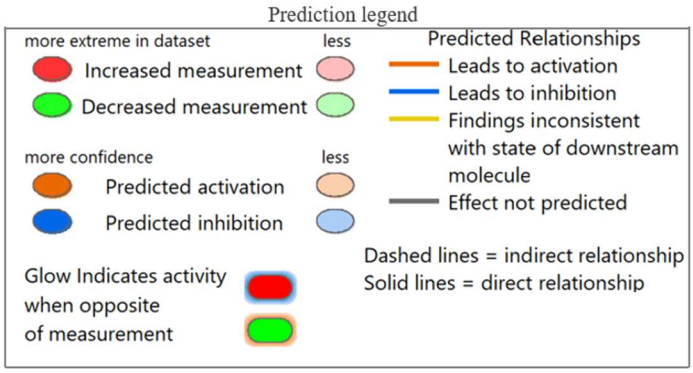
3.4.1. Classification applies to Regressor model.

Regressor model predicted values from the selected models were apply to a classification approach to determine whether these values are into low BCC proliferation rate < 1.5 or high proliferation, in order to make the evaluation of the accuracy a ROC curve is performed to visualize its specificity and sensitivity with a confusion matrix for the total number of observations.

3.5. Enrichment and Pathway analysis of the differential expressed protein

Associate DEPs identified in steps 3.3 and 3.4 that significantly contributed to understand obesity and cancer relationship. A table with DEPs gene name information is extracted including differential information like Mean Diff, *P-value* and FDR. Molecular function and cellular component are analyze using IPA QIAGEN dataset (https://www.qiagen.com) leaving species by default *Homo Sapiens* and performing a Core analysis against the knowledge base, canonical pathways and networks selection criteria is base is Z-Score and the results offer information under color codes shows in **Table 1**.

Table 1. Color codes for canonical pathways and network figure representations take from IPA.



4. RESULTS

4.1. Annotation and clustering

To explore the proteomic patterns embedded within breast adipose tissue-derived EVs, an extensive and comparative preprocessing strategy was first applied. Peptides with less than 20% sequence coverage were excluded. Intensity values were log2-transformed, missing values imputed using Iterative Imputer, and data standardized via StandardScaler(). These steps achieved normal intensity distributions across all samples. For dimensionality reduction techniques including PCA and UMAP were compared. UMAP, particularly with optimized parameters (n_neighbors=3, min_dist=0.001, metric='cosine'), outperformed PCA, demonstrating superior sample separation. Parameter tuning was conducted using a grid search (n_neighbors = [3, 4, 5, 10, 15, 30], min_dist = [0.000, 0.001, 0.005, 0.010, 0.050], metrics = ['euclidean', 'cosine', 'manhattan']) **Figure 4.**

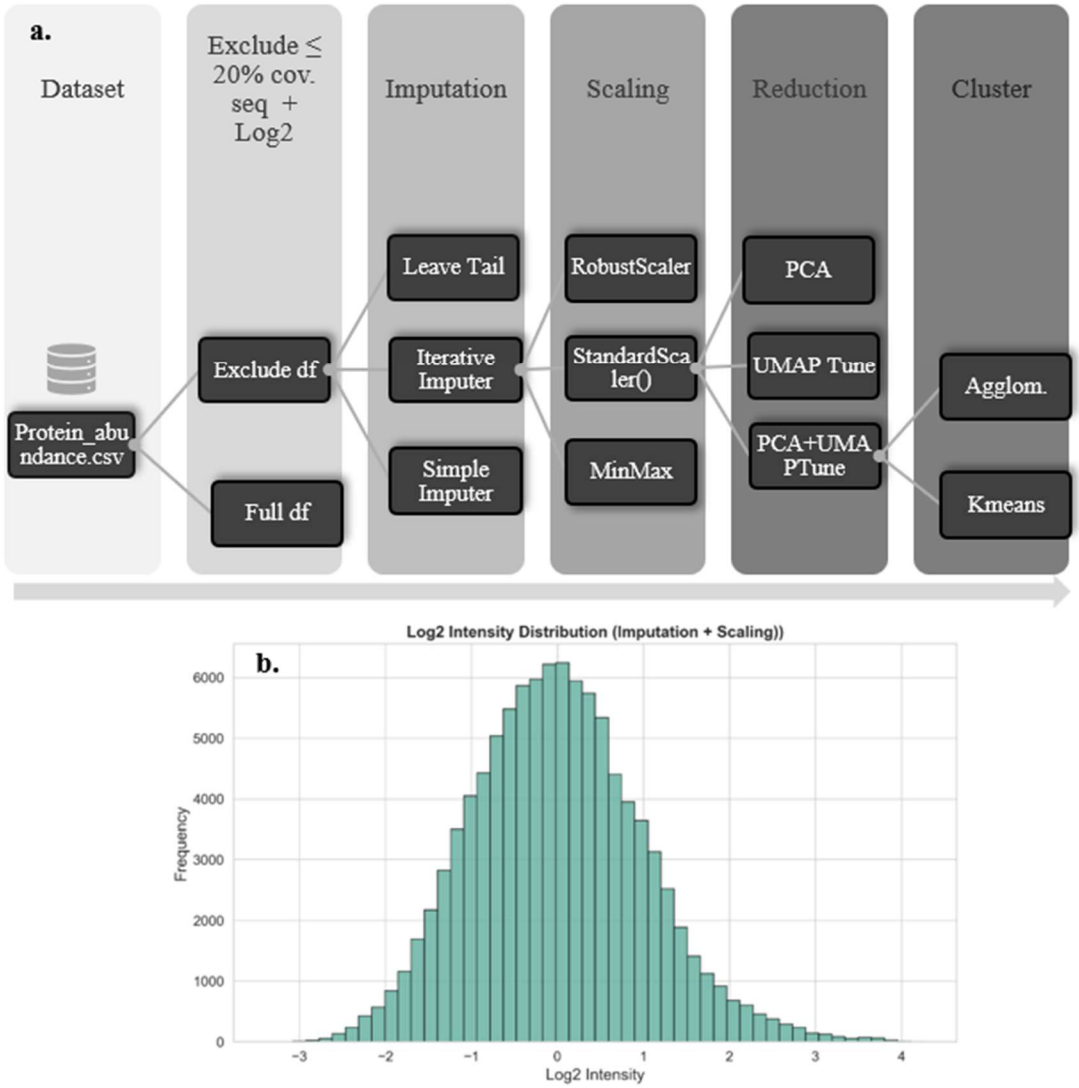


Figure 4. Summarization of optimal strategy selection for clustering. a. Final selection of all the comparison techniques (Exclude low quality peptides, Iterative Imputer, StandardScaler(), PCA+UMAP

tune and K-means). **b.** Distribution of log2-transformed intensity values after Iterative Imputer and StandardScaler() processing.

4.2. K-Means as better clustering approach

K-means clustering following dimensionality reduction revealed clear stratification into two main sample clusters, independent of clinical metadata such as for BMI ($p = 0.78$) and BCC ($p = 0.47$) proliferation index, T-tests showed no statistical significance, suggesting that protein expression patterns, rather than patient phenotypes, drive the clustering behavior. Implying the existence of distinct proteomic signatures among EVs that may reflect underlying biological mechanisms not captured by traditional clinical variables. The cluster samples classification includes 20 individuals against 28 in cluster 2 that shows a spare distributiondepicted in **Figure 5.**

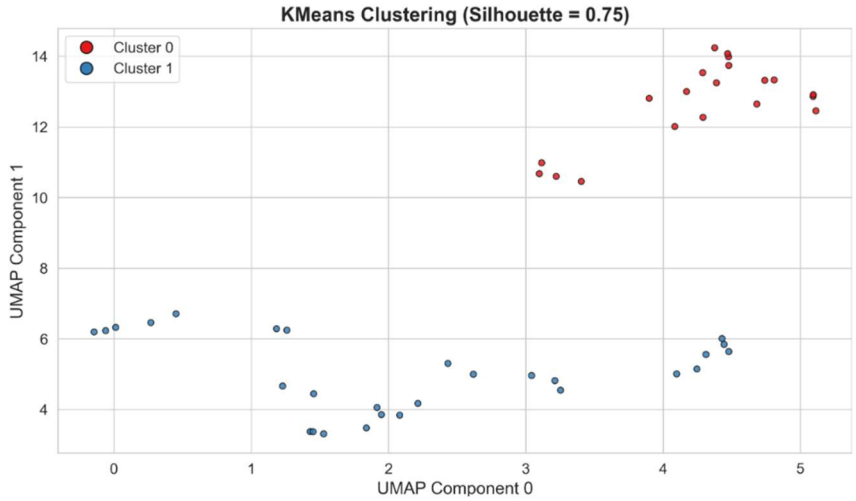


Figure 5. K-Means cluster after UMAP dimensional reduction. expression values after dimensionally reduction leads to distinct cluster formation into 2 groups, each point in the plot signifies a sample classify into cluster 0 (red) and cluster 1 (blue), expression values after dimensionally reduction leads to distinct cluster formation

4.3. Over 600 differentially expressed proteins (DEPs) identified among clusters

DEA was performed using both a standardized pipeline (AlphaPepStats) and a custom-built pipeline. Proteins were filtered based on significance thresholds using T-test statistical comparison metric ($p < 0.005$ and $|\log_2 FC| > [0.25 - 2]$), yielding distinct sets of upregulated and downregulated proteins across the two identified clusters. The AlphaPepStats pipeline was only compatible with the MaxQuant format, therefore the selected pre-processing was carried out by extracting intensity values, modifying and substituting them into the original table, pipeline interface only offers [1 - 2] threshold which makes the comparison limited. The clustering structure revealed relatively short distances between groups, possibly reflecting < 80% shared information (SI) or overlapping variance. Given these limitations, further analysis proceeded exclusively with the custom pipeline, which allowed tuning of fold change thresholds to lower values ($|\log_2 FC| \geq 0.25–0.5$). This more sensitive threshold captured the most relevant

differences between clusters, as illustrated in **Figure 6**. Our FC threshold is selected according the wide biological relevance, in order to explore and look potential shifts across clusters based only on their protein expression profile, we seek to capture small alterations that can be interpretate as effectors downstream for the following analysis.

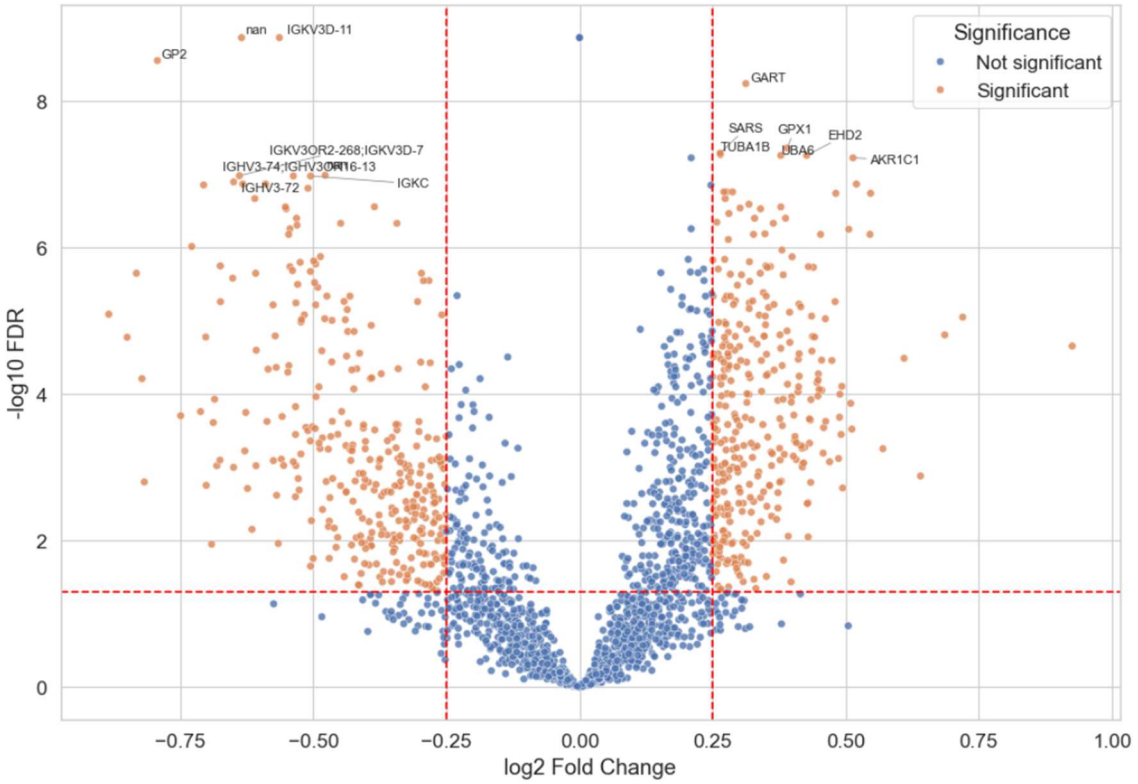


Figure 6. Top 10 most representative DEPs for FC [0.25 - 0.50]. Annotated volcano plot showing expression distribution of Top 10 up/down regulated DEPs. FC 0.50 keeps a reduced list of the shorter distance identification. Distance based on average mean values between clusters.

A Total of 665 DEPs were identified with 333 upregulated in Cluster 1 against 332 downregulated in cluster 2.

4.3.1. PPI shows more interactions nodes than expected

Protein-protein interaction networks constructed via the STRING database highlighted several densely connected subnetworks with an PPI enrichment p-value < 1.0e-16 that shows almost 3 times more edges than expected, supporting the functional coherence of differentially expressed proteins. TSV table from STRING was extracted and filtered by High Confidence > 0.7 interaction degree to keep only those highly related each other. For cluster 1, two distinct dense interactions were found, leaving behind 94 DEPs that does not present a strong interaction among cluster items. Similarly, the analysis was performed by a Fold change FC = 0.50, reducing the list to 13 DEPs, were only 3 interaction were identified (AKR1C1 and AKR1C2) without showing any significant interaction, different from cluster 2 that still keeps significant interaction with both FC values, even after excluding low confidences interaction. For FC = 0.25 just one dense cluster was visualize and for 0.50 remains 15 of them with significancy, (A2M, MMP2,

LTF, SERPINF1, CP, KLKB1, MMP9, SERPINA3, AZU1, SERPINA4, CAMP, COL14A1, LRG1, MMP14, and RECK).

4.4. Predictive Modeling of Cell Proliferation Behavior

To evaluate the potential of EV-derived proteomic profiles in predicting biological outcomes, a regression-based machine learning approach was implemented using 39 samples with available BCC proliferation indixes without considering DEA cluster result that does not identified any correlation with the metadata (BMI, BCC). Among several algorithms tested (Random Forest, Logistic Regression, Decision Tree, and KNN), the Decision Tree model exhibited the best predictive performance, with an R^2 value of 0.87, RMSE of 0.17, and MAE of 0.14. A comprehensive parameter grid search was conducted for all models to ensure optimal performance,shown in **Table 2.**

Table 2. Metrics models comparison. Performance metrics for predicting BCC proliferation index across tested algorithms.

Models		R^2	RMSE	MAE
Random Forest	-	0.62	0.3	0.22
Logistic Regression	-	0.61	0.3	0.22
Decision Tree	-	0.87	0.17	0.14
KNN	-	0.2	0.43	0.31

Given the limited sample size, the model's predictive structure remains robust and suitable for future validation. Additionally, a simulated classification of BCC proliferation was visualized using the continuous predictions generated by the regression models. Feature importance analysis from the DT model enabled the identification of 10 protein predictors contributing to cell proliferation behavior, STRING analysis revealed no statistically significant interactions among them. Despite this, their individual predictive contributions were substantial, shown in **Figure 7**. These proteins were further examined for biological relevance by GeneCart, with results summarized in **Table 3**.

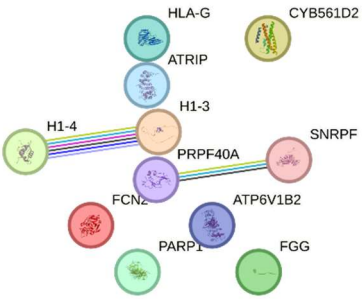


Figure 7. 10 features representative with no significant interaction. Most representative genes names identified by the regressor model predicted values. Gene name represented in every node with no significant interaction between them, except for PRPF40A and SNRPF.

Table 3. Pathways describe in GeneCart for every DEPs of DT model and their importance weight

Pathways	Importance	Genes
Immunosupression and aggression	0.513	CYB561D2
Insulin receptor recycling and MITF-M-dependent gene expression	0.20	ATP6V1B2
Processing of Capped Intron-Containing Pre-mRNA and Processing of Capped Intronless Pre-mRNA.	0.131	SNRPF
Apoptosis and survival FAS signaling cascades and Transcription-Coupled Nucleotide Excision Repair (TC-NER)	0.051	PARP1
Initial triggering of complement and Complement cascade.	0.040	FCN2
Cytokine Signaling in Immune system	0.026	HLA-G
Cellular responses to stimuli and Programmed Cell Death	0.016	H1-3; H1-4
Processing of Capped Intron-Containing Pre-mRNA.	0.009	PRPF40A
Signaling downstream of RAS mutants and Toll Like Receptor 7/8 (TLR7/8) Cascade	0.009	FGG
Homologous DNA Pairing and Strand Exchange and HDR through Homologous	0.001	ATRIP

The predicted values in a range of [0.4 - 2.8] were correctly predicted and compare in a hybrid classification model that include classification into low and high cell proliferation index threshold set 1.5 BCC proliferation index giving an AUC of 0.98 classifying correctly the predicted values that can build the base for a more robust model that can be trained by similar data sets, the results are summarized in **Figure 8.** The 1,5 BCC proliferation index threshold is based in previous experience of the group, reflecting biological behavior of MCF7 cells treated with O-OW EVs.

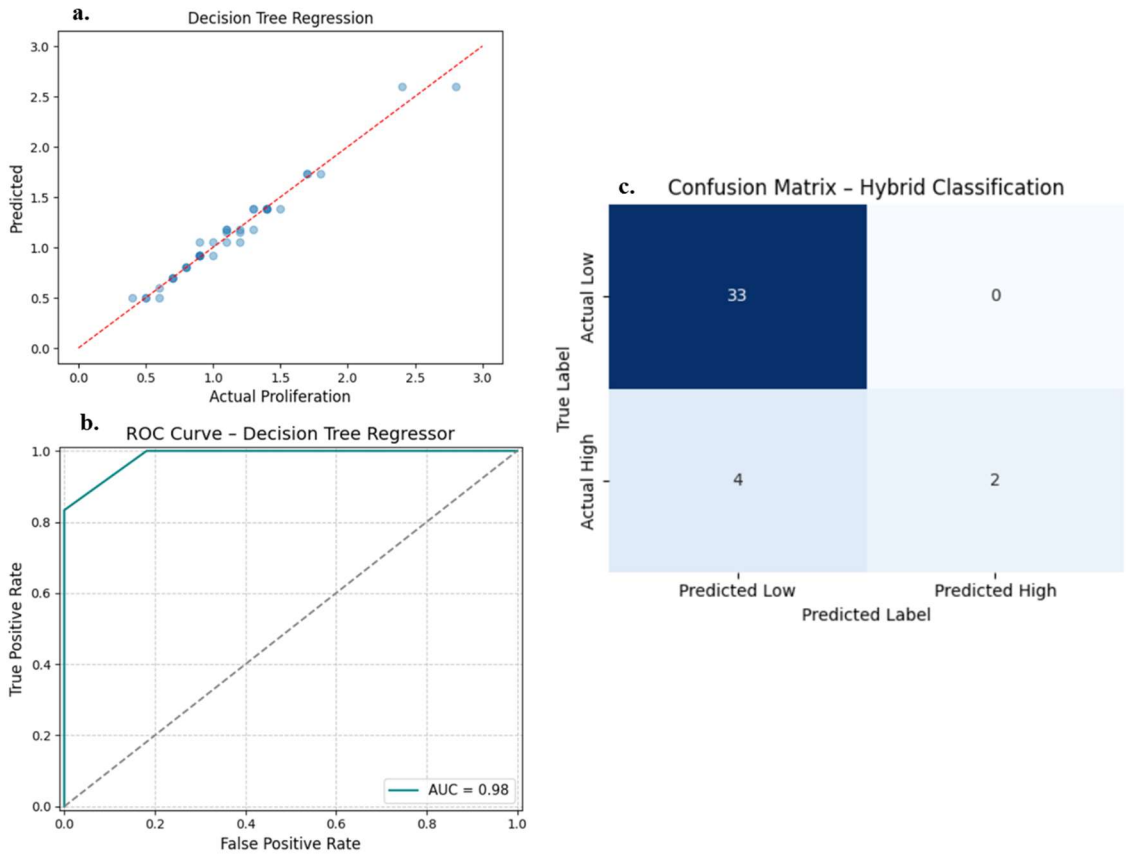


Figure 8. Hybrid classification model using predicted proliferation index. a. blue dots represent the predicted BCC proliferation index values that adjust perfectly to real values. **b.** Regressor model apply to classification into high and low proliferation shows AUC > 0.90 generating what is shown in figure **c.** Confusion matrix well classify TP and TN for the total of samples.

4.5. Canonical pathways related to obesity and cancer identified in core analysis

To interpret the potential biological relevance of the DEPs a core pathway analysis was performed using IPA focus in DEPs (using gene names) with high-confidence (HC > 0.7) interactions, as determined by STRING. Core pathway analysis identifies the most significantly enriched canonical pathways based on a given set of differentially expressed genes or proteins. It helps uncover underlying biological mechanisms, signaling cascades, and functional relationships within the data. Canonical pathways were ranked according to statistical significance and biological relevance, we exclude EVs related pathways whose enrichment would be inherently overrepresented in the canonical pathway list due to the EV origin of the samples without provide any new insight about the interaction of their cargo with the ATME. Instead, we focus on obesity and cancer-related mechanisms. The dataset was organized by Mean Diff, *p-value*, and FDR, highlighting the importance of pathways related with tumor cell lines development, sensitivity or colony formation as the main cores identified by IPA shown in **Figure 9**. Top-ranked proteins are shown in **Table 4**. To main canonical pathways are introduce in this study as the most significant for EVs cargo contribution to cancer (Estrogen-dependent breast cancer signaling and tumor microenvironment) . Interestingly the DEPs AKR1C1 and AKR1C2 associated to cluster 1 with a FC = 0.5 now shows a second layer interaction due the bulk of DEPs upload in the core analysis, mainly for ER regulation; belonging to Aldo-keto reductase superfamily and associated with lipid metabolism, previously identified as target for new hormone-based therapy strategies in primary BC (Wenners et al., 2016), some other are specifically aligned in IPA to the DEP set at FC = 0.5 such as MMP9 and MMP14, both metalloproteinases playing a role in tumor growth (Têtu et al., 2006; Vos et al., 2021)

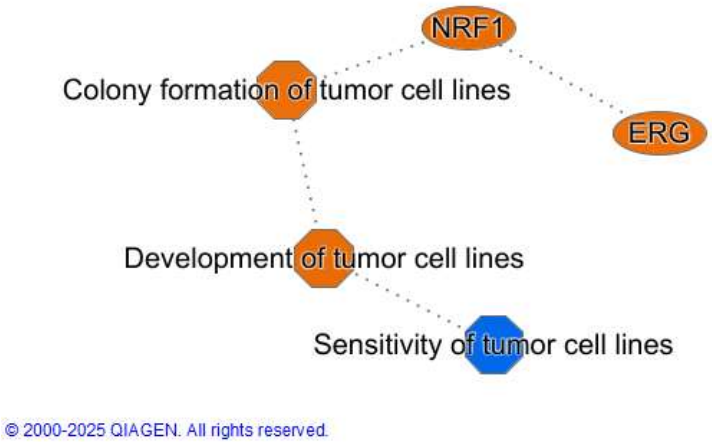


Figure 9. Graphical summary for global canonical pathways.

Estrogen-dependent breast cancer signaling is a molecular pathway that describes how ER, particularly estradiol (E2), promotes the development and progression of hormone-responsive breast cancer through estrogen receptor-mediated mechanisms.
This pathway was significantly enriched (Z-score = 2.18; 23 mapped molecules),

highlighting the central role of E2 in promoting tumorigenesis via both genomic and non-genomic actions mediated by estrogen receptor alpha (ESR1). The pathway shows interactions involving IGF1R, EGFR, ERK1/2, and downstream effectors such as STAT3, FOS, ELK1, and CREB, which are related to cell cycle progression and proliferation **Figure 10**. These signaling cascades reflect hormone-dependent mechanisms often enhanced in ER+ BC subtypes.

Table 4. Top analysis molecules identified by IPA. Mean difference represent the mean absolute between cluster 1 vs cluster 2 per each DEP. Arrows direction denotate upregulation ↑ and dowregulation ↓

CLUSTER 1 VS CLUSTER 2					
Gene		**Mean Diff**	**p-value**	**FDR**	
PLIN5	-	0.9257	2.0E-06	2.2E-05	
NAT1	-	0.7200	6.0E-07	9.0E-06	
KLHL31	-	0.6862	1.2E-06	1.6E-05	
SDSL	-	0.6407	3.3E-04	1.3E-03	
KLC4	-	0.6097	3.3E-06	3.3E-05	↑
ZC3HAV1L	-	0.5701	1.1E-04	5.6E-04	
AKR1C2	-	0.5468	2.5E-09	1.8E-07	
MAP2K6	-	0.5457	1.8E-08	6.7E-07	
NEK7	-	0.5203	1.3E-09	1.4E-07	
AKR1C1	-	0.5138	3.7E-10	6.0E-08	
IGLV5-45	-	-0.8845	5.2E-07	8.2E-06	
IGHV3-49	-	-0.8499	1.4E-06	1.7E-05	
SPRR3	-	-0.8327	9.0E-08	2.3E-06	
MMP9	-	-0.8222	7.8E-06	6.2E-05	
FCN2	-	-0.8178	4.2E-04	1.6E-03	
GP2*	-	-0.6815	2.3E-04	9.6E-04	↓
CTHRC1	-	-0.7494	3.1E-05	2.0E-04	
IGKV3D-20	-	-0.7286	2.7E-08	9.7E-07	
MMP2	-	-0.7121	2.6E-05	1.7E-04	
COL14A1	-	-0.7023	1.4E-06	1.7E-05	

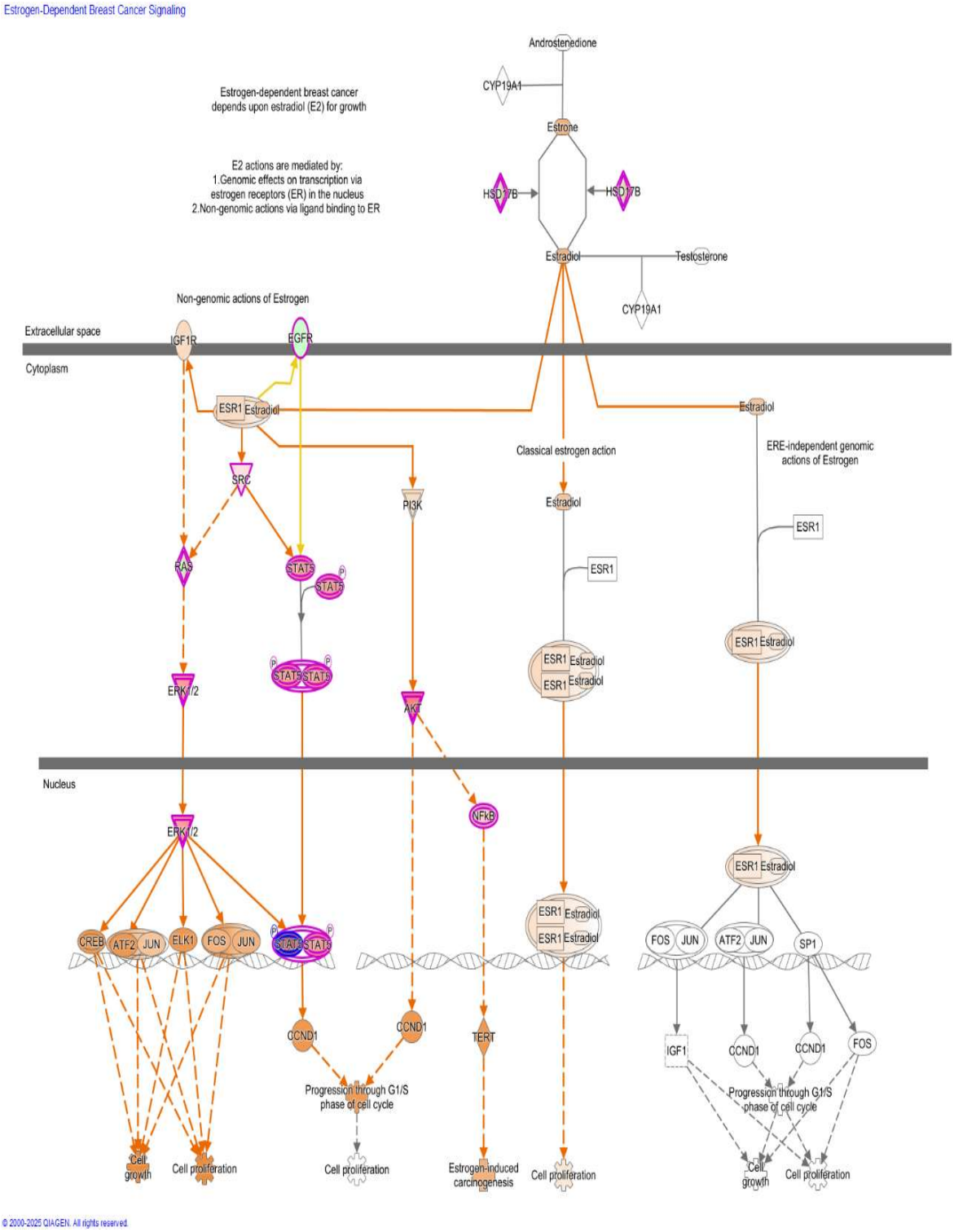


Figure 10. IPA visualization of the Estrogen-Dependent Breast Cancer Signaling pathway. Highlighted nodes indicates molecules detected in the dataset, include estradiol (E2), ESR1, PI3K, ERK1/2 , and transcription factors associated with cell proliferation. For orange nodes the molecules predicted to be activates, pink nodes proteins differentially expressed, gray nodes not present in dataset, solid orange arrows direct activation while dashed orange arrows represent indirect activation.

Tumor microenvironment (TME) refers to the complex and dynamic network of non-cancerous cells and extracellular components that interact with tumor cells, influencing cancer progression. The second significantly enriched pathways (Z-score = 1.76; 32

mapped molecules) highlights significant interactions within the tumor microenvironment, including adipocytes, immune cells, endothelial cells, and extracellular matrix (ECM) remodeling factors. DEPs were involved in processes like angiogenesis, immune suppression via PD-L1 and HLA-G, VEGF signaling, and matrix degradation by MMP9 and MMP14, contributing to tumor support **Figure 11**. These results indicates that the cargo from EVs derived from obese adipose tissue may promote a pro-tumorigenic microenvironment by influencing both stromal remodeling and immune regulation, according the knowledge base from IPA.

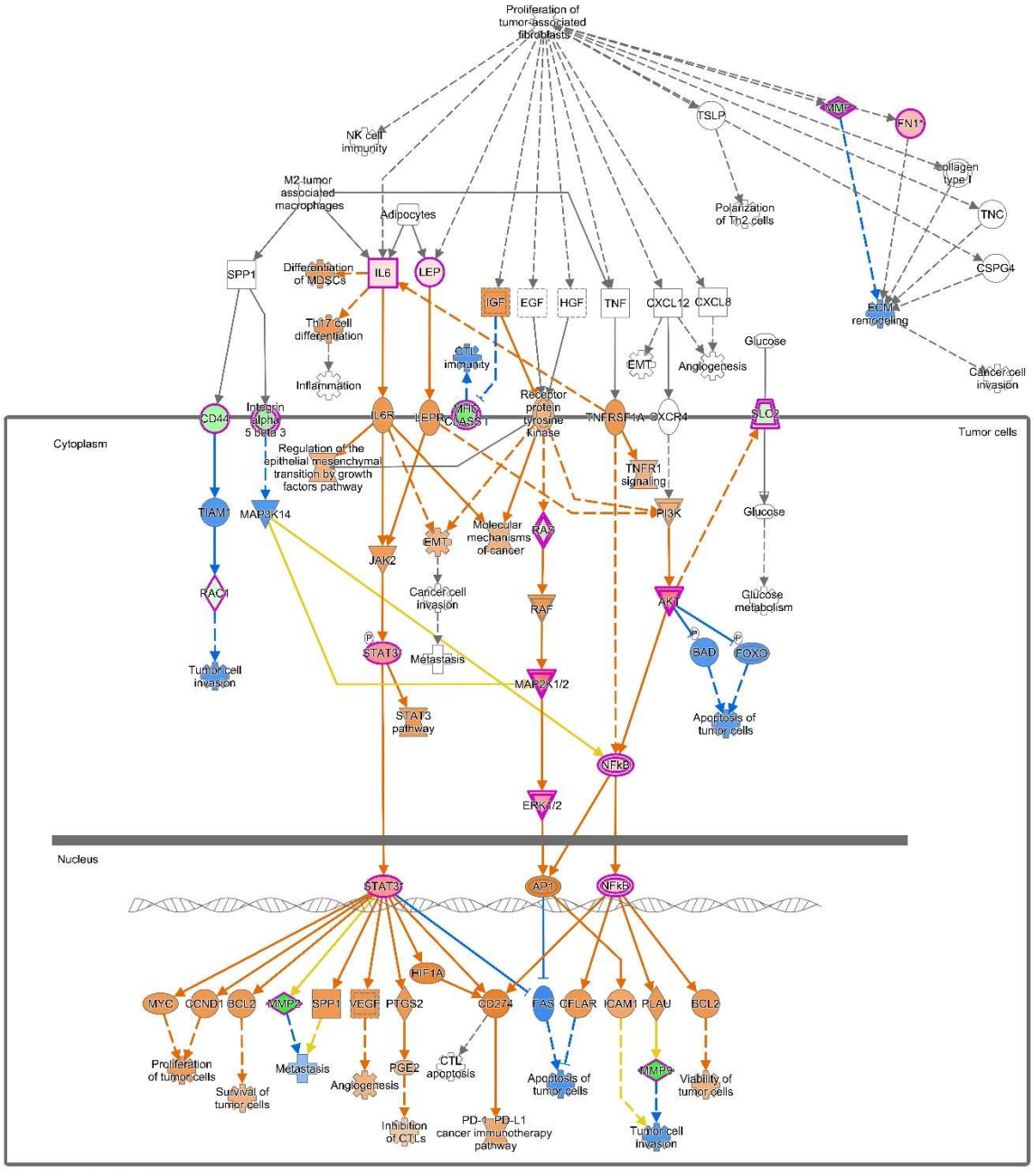


Figure 11. IPA visualization of the Tumor Microenvironment pathway. Adapted figure from IPA showing the most representative DEPs mapped in the canonical pathway. Proteins implicated in immune modulation, ECM remodeling, and angiogenesis are represented, supporting the role of EVs in shaping

tumor supportive conditions. Color code follows the same pattern than *figure 10* blue nodes and arrows indicate inhibition.

4.5.2. β-catenin network associated function EV cargo to cancer pathways

The *β*-catenin (CTNNB1) signaling network plays a key role in cancer by regulating cell proliferation, invasion, and tissue homeostasis, and is frequently dysregulated in tumor development. Among the top interaction networks identified through IPA core analysis different from canonical pathway identification, network construction does not include pathways related information explicitly, instead it looks internal interaction from the DEP list (Krämer et al., 2014). Network 2 is the most biologically relevant, with a score of 30 and annotated functional associations with cancer, gastrointestinal disease, and hepatic system disease. This network is centered on *β*-catenin, frequently dysregulated in cancer and closely linked to tumor cell proliferation, invasion.
Several proteins from DEPs dataset mapped directly to this network, including GPC1 a known EVs marker previously associated with cancer progression (Lucien et al., 2019), similarly SERPINA4, and APOD, contribute to extracellular matrix regulation and lipid metabolism the interactions are shown in **Figure 12.** Notably align to TME canonical pathway, suggesting these DEPs as important modulators of tumor signaling.

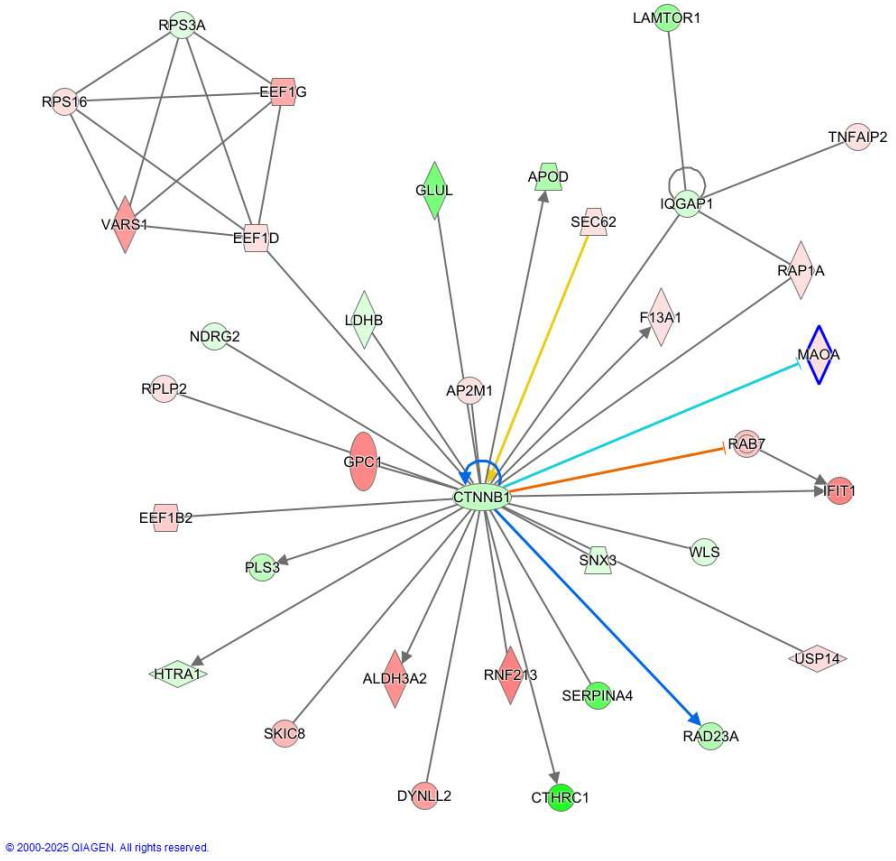


Figure 12. CTNNB1-Centered Protein Interaction Network Identified in EVs Derived DEPs. High confidence interaction with molecules participating in cell proliferation and adhesion by CTNNB1 as central node and inhibited regulator. Green nodes proteins predicted to be downregulated in the dataset, red nodes, proteins predicted to be upregulated, white nodes are no DE; shapes indicate molecular classes, ovals: enzymes, diamonds: Transcription regulators, rounded rectangles: transporter, vertical ellipses:

cytokines or signaling proteins and Hexagons: others.

Other high score networks were identified by IPA using the knowledge base that related the EVs cargo with STAT5A node linking to immune signaling and connect with ER receptor pathways via RXR and POR not shown in this study.

5. DISCUSSION

In the current study, we integrated exploratory data science with biological interpretation to investigate how proteomic signatures derived from breast adipose tissue EVs may influence cancer-associated processes, particularly in the context of obesity. We combined unsupervised clustering, differential expression analysis (DEA), predictive modeling of proliferation, and IPA-based pathway enrichment to delineate a potential biological system according the networks.

The most important step is the preprocessing that began with a dimensionality reduction and unsupervised clustering strategy to uncover hidden proteomic patterns; omics data implies high dimensions that are traditionally reduce by linear reductions algorithms such as PCA or nonlinear as t-SNE or UMAP, however mix approach have been contemplated by feature selection, making a double distance reduction that preserve the information in a low dimensional projections, these arrives because the limitation of the nonlinear reduction that preserve locally, but struggles to capture the global patterns , (Pal & Sharma, 2020; Wenskovitch et al., 2018). Our results confirm this advantage, as the combination exposed a subtle but biologically meaningful separation into two main clusters, despite a Silhouette Index > 70% this distance across clusters with statistical meaningful structure reflect an expression shift that is not driven by clinical metadata, instead it detected molecular differences.

Previous works for model prediction designed based on open source dataset extensively used like Wisconsin Diagnostic Breast Cancer (WDBC) and Breast Cancer Coimbra Dataset (BCCD) rely heavily on metadata and are trained for binary diagnosis (Rasool et al., 2022) or long term data collection in hospitals (Rabiei et al., 2022) the inclusion of only clinical information related to BC factors like age, BMI and analytical metrics and mammography offers limit information about relationships and deeper understanding about what molecular features may influences in this factors. Furthermore, similar approaches are used for hyperparameter optimization and comparing predictive models in which they coincide RF, LR and KNN, with accuracy of 98.06% for LR. Every technique is used to be adapted and adjust to every dataset, and a bigger input helps in the reliability of the model, in our case the limited sample size (n=39 for BCC) and lack of a separate test set hindered model validation. Moreover, the small number of features with significant importance scores raises concerns about overfitting and model fragility in unseen data.

To address interpretability and improve diagnostic value we implemented a hybrid model that converts regression output into binary classification strategy used previously for logistic regression model (Guo et al., 2006), this approach is used to represents low and high proliferation potential according the protein expression profile that may influence

the dynamic behavior under the influence of O-EVs.

The most significant pathways identified were ER dependent BC Signaling and the TME. Both are well established in obesity linked breast cancer where O-EVs identified DEPs in cluster analysis are active modulators of tumor biology.

The IPA identified pathways ER dependent BC Signaling, TME, and the β-Catenin interaction network collectively underscore the potential mechanisms by which EVs derived from obese adipose tissue may contribute to breast cancer progression. EV cargo appears to influence estrogen receptor signaling, enhancing proliferation in ER+ BC subtypes through both genomic and non-genomic actions. Simultaneously, EV associated proteins impact the tumor microenvironment by promoting angiogenesis, immune suppression, and matrix remodeling—features commonly linked to tumor support and aggressiveness. Notably, the enrichment of β-catenin-centered signaling networks, including components involved in extracellular matrix regulation, proteostasis, and lipid metabolism, further supports a role for obesity associated EVs in modulating oncogenic pathways. Together, these findings suggest that adipose-derived EVs may act as mediators of obesity driven tumorigenesis by coordinating hormone signaling, immune evasion, and stromal remodeling in BC.

Other studies have similarly highlighted genes such as MMP9, GYG2, PRKAR2B as hub molecules in obese BC patients uncover by a FC < 0.67 (Comertpay & Gov, 2022) which appears within our canonical pathways; it has been reported that MMP9 describe to promote immunological TME in BC progression (Reggiani et al., 2017). In other studies obesity and BC have been link the role of ER in BC by pointing aromatases as a main interaction node regulated by adipocytes (Gérard & Brown, 2018) and proteins like CLEC3B previously identified as a strong predictor in cancer models also appeared in our Cluster 1 and were involved in the TME pathway (Kallah-Dagadu et al., 2025). More specifically recent ML model based on tumor EV protein profile have been develop to predicted the potential of EVs for cell invasion and proliferation, however excluding important factor that may include this as obesity (Bukva et al., 2023). Our finding also reveals strong overlap between high BMI associated proteins reported in the original study by Liu et al. (2023) (Liu et al., 2023), where we consistently find SERPINF2, SERPINA1, SERPINA4, CA1, PROS1, ITGA2B, ITGB3, GP9 in cluster 2, DYNC1LI2, AKR1C3, CAV1, CAV2, UTRN, KIF5B For the upregulates, consistently in cluster 1. While some of the DT model's most predictive features were not directly found in the enriched pathways their pathways participations are highly related to immune response and can work as a regulators effectors that needs further analysis. Differently in other studies in micro vesicles an extracellular metalloproteinase inducer (EMMPRIN) is identified as a marker for BC this effect is describe for not being mediated by matrix metalloproteinases, but by activation of the p38/MAPK signaling pathway in the tumor cells (Menck et al., 2015) in our results MMP9 and MMP14 extensively mentioned reveal a distinct mechanism within the EVs (exosome) associate to TME and β-catenin network Our results are exploratory, yet methodologically consistent with larger studies applying omics for cancer risk modeling. Even though our models are not robust for external testing, they offer a proof of concept that can be enhanced through future datasets and multi-omics integration. Hybrid models like the one we applied, when trained with a

larger cohort and additional covariates may allow real world deployment for early risk prediction. Future research should focus on combining protein expression with established factors such as age, BMI, BCC proliferation, and genomic mutations to build a robust predictive system.

6. CONCLUSIONS

Although the predictive models developed in this study are not yet strong enough for clinical application, they provide a starting point for the identification of molecular signatures involved in obesity-related breast cancer risk. The combination of exploratory clustering, hybrid modeling strategies, and IPA enrichment analyses revealed both familiar and potentially novel players in tumor development. With larger datasets, inclusion of external validation cohorts, and inclusion of additional clinical metadata, this approach can evolve into a powerful diagnostic tool. Furthermore, concordance of machine learning output with biologically validated pathways suggests that the exploratory findings capture disease-relevant mechanisms. These results are worthy of further investigation to add to our understanding of the relationship between cancer and obesity, according to IPA's knowledge base.

7. BIBLIOGRAPHY

Al Mudawi, N., & Alazeb, A. (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors*, *22*(11). https://doi.org/10.3390/s22114132

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). *Gene Ontology: tool for the unification of biology The Gene Ontology Consortium**. http://www.flybase.bio.indiana.edu

Avgerinos, K. I., Spyrou, N., Mantzoros, C. S., & Dalamaga, M. (2019). Obesity and cancer risk: Emerging biological mechanisms and perspectives. In *Metabolism: Clinical and Experimental* (Vol. 92, pp. 121–135). W.B. Saunders. https://doi.org/10.1016/j.metabol.2018.11.001

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–47. https://doi.org/10.1038/nbt.4314

Beckman, J. A., & Creager, M. A. (2016). Vascular complications of diabetes. In *Circulation Research* (Vol. 118, Issue 11, pp. 1771–1785). Lippincott Williams and Wilkins. https://doi.org/10.1161/CIRCRESAHA.115.306884

Bensmail, H., Golek, J., Moody, M. M., Semmes, J. O., & Haoudi, A. (2005). A novel approach for clustering proteomics data using Bayesian fast Fourier transform. *Bioinformatics*, *21*(10), 2210–2224. https://doi.org/10.1093/bioinformatics/bti383

Bowers, L. W., Rossi, E. L., O'Flanagan, C. H., De Graffenried, L. A., & Hursting, S. D. (2015). The role of the insulin/IGF system in cancer: Lessons learned from clinical trials and the energy balance-cancer link. In *Frontiers in Endocrinology* (Vol. 6, Issue MAY). Frontiers Media S.A. https://doi.org/10.3389/fendo.2015.00077

Bukva, M., Dobra, G., Gyukity-Sebestyen, E., Boroczky, T., Korsos, M. M., Meckes, D.

G., Horvath, P., Buzas, K., & Harmati, M. (2023). Machine learning-based analysis of cancer cell-derived vesicular proteins revealed significant tumor-specificity and predictive potential of extracellular vesicles for cell invasion and proliferation – A meta-analysis. *Cell Communication and Signaling*, *21*(1). https://doi.org/10.1186/s12964-023-01344-5

Cairns, R. A., Harris, I. S., & Mak, T. W. (2011). Regulation of cancer cell metabolism. In *Nature Reviews Cancer* (Vol. 11, Issue 2, pp. 85–95). Nature Publishing Group. https://doi.org/10.1038/nrc2981

Callahan, N., Tullman, J., Kelman, Z., & Marino, J. (2020). Strategies for Development of a Next-Generation Protein Sequencing Platform. In *Trends in Biochemical Sciences* (Vol. 45, Issue 1, pp. 76–89). Elsevier Ltd. https://doi.org/10.1016/j.tibs.2019.09.005

Chen, L., Chu, C., Lu, J., Kong, X., Huang, T., & Cai, Y. D. (2015). Gene ontology and KEGG pathway enrichment analysis of a drug target-based classification system. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0126492

Chooi, Y. C., Ding, C., & Magkos, F. (2019). The epidemiology of obesity. *Metabolism: Clinical and Experimental*, *92*, 6–10. https://doi.org/10.1016/j.metabol.2018.09.005

Comertpay, B., & Gov, E. (2022). Identification of molecular signatures and pathways of obese breast cancer gene expression data by a machine learning algorithm. *Journal of Translational Genetics and Genomics*, *6*(1), 84–94. https://doi.org/10.20517/jtgg.2021.44

Conte, L., Rizzo, E., Civino, E., Tarantino, P., De Nunzio, G., & De Matteis, E. (2024). Enhancing Breast Cancer Risk Prediction with Machine Learning: Integrating BMI, Smoking Habits, Hormonal Dynamics, and BRCA Gene Mutations—A Game-Changer Compared to Traditional Statistical Models? *Applied Sciences (Switzerland)*, *14*(18). https://doi.org/10.3390/app14188474

Eichelser, C., Stückrath, I., Müller, V., Milde-Langosch, K., Wikman, H., Pantel, K., & Schwarzenbach, H. (2014). *Increased serum levels of circulating exosomal microRNA-373 in receptor-negative breast cancer patients* (Vol. 5, Issue 20). www.impactjournals.com/oncotarget

Feng, Q., Zhang, C., Lum, D., Druso, J. E., Blank, B., Wilson, K. F., Welm, A., Antonyak, M. A., & Cerione, R. A. (2017). A class of extracellular vesicles from breast cancer cells activates VEGF receptors and tumour angiogenesis. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14450

Ferreira, A. P. de S., Szwarcwald, C. L., Damacena, G. N., & de Souza Júnior, P. R. B. (2021). Increasing trends in obesity prevalence from 2013 to 2019 and associated factors in Brazil. *Revista Brasileira de Epidemiologia*, *24*. https://doi.org/10.1590/1980-549720210009.SUPL.2

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0801-4

Gérard, C., & Brown, K. A. (2018). Obesity and breast cancer – Role of estrogens and the molecular underpinnings of aromatase regulation in breast adipose tissue. In *Molecular and Cellular Endocrinology* (Vol. 466, pp. 15–30). Elsevier Ireland Ltd. https://doi.org/10.1016/j.mce.2017.09.014

Guo, H.-M., Lotus Shyu, Y.-I., & Chang, H.-K. (2006). *Combining Logistic Regression with Classification and Regression Tree to Predict Quality of Care in a Home Health Nursing Data Set*.

Hilario, M., & Kalousis, A. (2008). Approaches to dimensionality reduction in proteomic

biomarker studies. In *Briefings in Bioinformatics* (Vol. 9, Issue 2, pp. 102–118). https://doi.org/10.1093/bib/bbn005

Holger Husi, & SC Nat. (2019). *Computational Biology*. https://doi.org/http://dx.doi.org/10.15586/computationalbiology.2019

Hong, R., & Xu, B. (2022). Breast cancer: an up-to-date review and future perspectives. In *Cancer Communications* (Vol. 42, Issue 10, pp. 913–936). John Wiley and Sons Inc. https://doi.org/10.1002/cac2.12358

James, F. R., Wootton, S., Jackson, A., Wiseman, M., Copson, E. R., & Cutress, R. I. (2015). Obesity in breast cancer - What is the risk factor? In *European Journal of Cancer* (Vol. 51, Issue 6, pp. 705–720). Elsevier Ltd. https://doi.org/10.1016/j.ejca.2015.01.057

Kallah-Dagadu, G., Mohammed, M., Nasejje, J. B., Mchunu, N. N., Twabi, H. S., Batidzirai, J. M., Singini, G. C., Nevhungoni, P., & Maposa, I. (2025). Breast cancer prediction based on gene expression data using interpretable machine learning techniques. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-85323-5

König, L., Kasimir-Bauer, S., Bittner, A. K., Hoffmann, O., Wagner, B., Santos Manvailer, L. F., Kimmig, R., Horn, P. A., & Rebmann, V. (2018). Elevated levels of extracellular vesicles are associated with therapy failure and disease progression in breast cancer patients undergoing neoadjuvant chemotherapy. *OncoImmunology*, *7*(1). https://doi.org/10.1080/2162402X.2017.1376153

Kothari, C., Diorio, C., & Durocher, F. (2020). The importance of breast adipose tissue in breast cancer. In *International Journal of Molecular Sciences* (Vol. 21, Issue 16, pp. 1–33). MDPI AG. https://doi.org/10.3390/ijms21165760

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. In *Computational and Structural Biotechnology Journal* (Vol. 13, pp. 8–17). Elsevier B.V. https://doi.org/10.1016/j.csbj.2014.11.005

Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, *30*(4), 523–530. https://doi.org/10.1093/bioinformatics/btt703

Kumar, M. A., Baba, S. K., Sadida, H. Q., Marzooqi, S. Al, Jerobin, J., Altemani, F. H., Algehainy, N., Alanazi, M. A., Abou-Samra, A. B., Kumar, R., Al-Shabeeb Akil, A. S., Macha, M. A., Mir, R., & Bhat, A. A. (2024). Extracellular vesicles as tools and targets in therapy for diseases. In *Signal Transduction and Targeted Therapy* (Vol. 9, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-024-01735-1

LG, A., & AT, E. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, *04*(02). https://doi.org/10.4172/2157-7420.1000124

Lilite Sadovska, Janis Eglitis, & Aija Line. (2015). *Extracellular Vesicles as Biomarkers and Therapeutic Targets in Breast Cancer*. https://doi.org/35:6379-6390(2015)

Liu, S., Benito-Martin, A., Pelissier Vatter, F. A., Hanif, S. Z., Liu, C., Bhardwaj, P., Sethupathy, P., Farghli, A. R., Piloco, P., Paik, P., Mushannen, M., Dong, X., Otterburn, D. M., Cohen, L., Bareja, R., Krumsiek, J., Cohen-Gould, L., Calto, S., Spector, J. A., … Brown, K. A. (2023). Breast adipose tissue-derived extracellular vesicles from obese women alter tumor cell metabolism. *EMBO Reports*, *24*(12). https://doi.org/10.15252/embr.202357339

Logozzi, M., De Milito, A., Lugini, L., Borghi, M., Calabrò, L., Spada, M., Perdicchio, M., Marino, M. L., Federici, C., Iessi, E., Brambilla, D., Venturi, G., Lozupone, F., Santinami, M., Huber, V., Maio, M., Rivoltini, L., & Fais, S. (2009). High levels of exosomes expressing CD63 and caveolin-1 in plasma of melanoma patients. *PLoS*

ONE, *4*(4). https://doi.org/10.1371/journal.pone.0005219

Lualdi, M., & Fasano, M. (2019). Statistical analysis of proteomics data: A review on feature selection. *Journal of Proteomics*, *198*, 18–26. https://doi.org/10.1016/j.jprot.2018.12.004

Lucien, F., Lac, V., Billadeau, D. D., Borgida, A., Gallinger, S., & Leong, H. S. (2019). *Glypican-1 and glycoprotein 2 bearing extracellular vesicles do not discern pancreatic cancer from benign pancreatic diseases*. www.oncotarget.com

Macinnis, R. J., English, D. R., Gertig, D. M., Hopper, J. L., & Giles, G. G. (2004). *Body Size and Composition and Risk of Postmenopausal Breast Cancer*. http://aacrjournals.org/cebp/article-pdf/13/12/2117/1939325/2117-2125.pdf

Marti, A., Calvo, C., & Martínez, A. (2021). Ultra-processed food consumption and obesity—a systematic review. In *Nutricion Hospitalaria* (Vol. 38, Issue 1, pp. 177–185). ARAN Ediciones S.A. https://doi.org/10.20960/nh.03151

Matilainen, J., Berg, V., Vaittinen, M., Impola, U., Mustonen, A. M., Männistö, V., Malinen, M., Luukkonen, V., Rosso, N., Turunen, T., Käkelä, P., Palmisano, S., Arasu, U. T., Sihvo, S. P., Aaltonen, N., Härkönen, K., Caddeo, A., Kaminska, D., Pajukanta, P., … Rilla, K. (2024). Increased secretion of adipocyte-derived extracellular vesicles is associated with adipose tissue inflammation and the mobilization of excess lipid in human obesity. *Journal of Translational Medicine*, *22*(1). https://doi.org/10.1186/s12967-024-05249-w

Matthiesen, R., & Bunkenborg, J. (2013). Introduction to mass spectrometry-based proteomics. In *Methods in Molecular Biology* (Vol. 1007, pp. 1–45). Humana Press Inc. https://doi.org/10.1007/978-1-62703-392-3_1

Medzihradszky, K. F., & Chalkley, R. J. (2015). Lessons in de novo peptide sequencing by tandem mass spectrometry. In *Mass Spectrometry Reviews* (Vol. 34, Issue 1, pp. 43–63). John Wiley and Sons Inc. https://doi.org/10.1002/mas.21406

Melo, S. A., Sugimoto, H., O'Connell, J. T., Kato, N., Villanueva, A., Vidal, A., Qiu, L., Vitkin, E., Perelman, L. T., Melo, C. A., Lucci, A., Ivan, C., Calin, G. A., & Kalluri, R. (2014). Cancer Exosomes Perform Cell-Independent MicroRNA Biogenesis and Promote Tumorigenesis. *Cancer Cell*, *26*(5), 707–721. https://doi.org/10.1016/j.ccell.2014.09.005

Menck, K., Scharf, C., Bleckmann, A., Dyck, L., Rost, U., Wenzel, D., Dhople, V. M., Siam, L., Pukrop, T., Binder, C., & Klemm, F. (2015). Tumor-derived microvesicles mediate human breast cancer invasion through differentially glycosylated EMMPRIN. *Journal of Molecular Cell Biology*, *7*(2), 143–153. https://doi.org/10.1093/jmcb/mju047

Meunier, B., Dumas, E., Piec, I., Béchet, D., Hébraud, M., & Hocquette, J. F. (2007). Assessment of hierarchical clustering methodologies for proteomic data mining. *Journal of Proteome Research*, *6*(1), 358–366. https://doi.org/10.1021/pr060343h

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., & Marks, J. S. (n.d.). *Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001*. http://jama.jamanetwork.com/

Neagu, A.-N., Whitham, D., Buonanno, E., Jenkins, A., Alexa-Stratulat, T., Tamba, B. I., & Darie, C. C. (2021). Proteomics and its applications in breast cancer. In *Am J Cancer Res* (Vol. 11, Issue 9). www.ajcr.us/

Nevadunsky, N. S., Van Arsdale, A., Strickler, H. D., Moadel, A., Kaur, G., Levitt, J., Girda, E., Goldfinger, M., Goldberg, G. L., & Einstein, M. H. (2014). Obesity and age at diagnosis of endometrial cancer. *Obstetrics and Gynecology*, *124*(2 PART1), 300–306. https://doi.org/10.1097/AOG.0000000000000381

Nwafor, J. N., Figueroa, A. S., Okobi, O. E., Ojukwu, G., Fanegan, E. J., Nyamekye-

Affel, R., Oyewole, B. O., Omotunde, O., Mamah, G. N., & Muoghalu, N. (2025). Obesity-Associated Cancers: A United States Cancer Statistics (USCS) Database Analysis. *Cureus*. https://doi.org/10.7759/cureus.84610

Pal, K., & Sharma, M. (2020). Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. *Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020*, 1106–1110. https://doi.org/10.1109/I-SMAC49090.2020.9243502

Park, S., Koo, J. S., Kim, M. S., Park, H. S., Lee, J. S., Lee, J. S., Kim, S. Il, & Park, B. W. (2012). Characteristics and outcomes according to molecular subtypes of breast cancer as classified by a panel of four biomarkers using immunohistochemistry. *Breast*, *21*(1), 50–57. https://doi.org/10.1016/j.breast.2011.07.008

Quail, D. F., & Dannenberg, A. J. (2019). The obese adipose tissue microenvironment in cancer development and progression. In *Nature Reviews Endocrinology* (Vol. 15, Issue 3, pp. 139–154). Nature Publishing Group. https://doi.org/10.1038/s41574-018-0126-x

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of Biomedical Physics and Engineering*, *12*(3), 297–308. https://doi.org/10.31661/jbpe.v0i0.2109-1403

Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. *International Journal of Environmental Research and Public Health*, *19*(6). https://doi.org/10.3390/ijerph19063211

Reggiani, F., Labanca, V., Mancuso, P., Rabascio, C., Talarico, G., Orecchioni, S., Manconi, A., & Bertolini, F. (2017). Adipose progenitor cell secretion of GM-CSF and MMP9 promotes a stromal and immunological microenvironment that supports breast cancer progression. *Cancer Research*, *77*(18), 5169–5182. https://doi.org/10.1158/0008-5472.CAN-17-0914

Rosen, E. D., & Spiegelman, B. M. (2014). What we talk about when we talk about fat. In *Cell* (Vol. 156, Issues 1–2, pp. 20–44). Elsevier B.V. https://doi.org/10.1016/j.cell.2013.12.012

Schmidt, A., Forne, I., & Imhof, A. (2014). Bioinformatic analysis of proteomics data. In *BMC systems biology* (Vol. 8, p. S3). https://doi.org/10.1186/1752-0509-8-S2-S3

Têtu, B., Brisson, J., Wang, C. S., Lapointe, H., Beaudry, G., Blanchette, C., & Trudel, D. (2006). The influence of MMP-14, TIMP-2 and MMP-2 expression on breast cancer prognosis. *Breast Cancer Research*, *8*(3). https://doi.org/10.1186/bcr1503

Vos, M. C., van der Wurff, A. A. M., van Kuppevelt, T. H., & Massuger, L. F. A. G. (2021). The role of MMP-14 in ovarian cancer: a systematic review. In *Journal of Ovarian Research* (Vol. 14, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13048-021-00852-7

Welsh, J. A., Goberdhan, D. C. I., O'Driscoll, L., Buzas, E. I., Blenkiron, C., Bussolati, B., Cai, H., Di Vizio, D., Driedonks, T. A. P., Erdbrügger, U., Falcon-Perez, J. M., Fu, Q. L., Hill, A. F., Lenassi, M., Lim, S. K., Mahoney, M. ỹ. G., Mohanty, S., Möller, A., Nieuwland, R., … Zubair, H. (2024). Minimal information for studies of extracellular vesicles (MISEV2023): From basic to advanced approaches. *Journal of Extracellular Vesicles*, *13*(2). https://doi.org/10.1002/jev2.12404

Wenners, A., Hartmann, F., Jochens, A., Roemer, A. M., Alkatout, I., Klapper, W., van Mackelenbergh, M., Mundhenke, C., Jonat, W., & Bauer, M. (2016). Stromal markers AKR1C1 and AKR1C2 are prognostic factors in primary human breast cancer. *International Journal of Clinical Oncology*, *21*(3), 548–556.

https://doi.org/10.1007/s10147-015-0924-2

Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., Leman, S., & North, C. (2018). Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 131–141. https://doi.org/10.1109/TVCG.2017.2745258

Wu, C., Dong, S., Huang, R., & Chen, X. (2023). Cancer-Associated Adipocytes and Breast Cancer: Intertwining in the Tumor Microenvironment and Challenges for Cancer Therapy. In *Cancers* (Vol. 15, Issue 3). MDPI. https://doi.org/10.3390/cancers15030726

Wu, Q., Li, B., Li, Z., Li, J., Sun, S., & Sun, S. (2019). Cancer-associated adipocytes: Key players in breast cancer progression. In *Journal of Hematology and Oncology* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13045-019-0778-6

Yaturu, S., & Jain, S. K. (2007). Obesity and type 2 diabetes. In *Obesity: Epidemiology, Pathophysiology, and Prevention* (pp. 139–154). CRC Press. https://doi.org/10.4236/jdm.2011.14012

Yoon, K. H., Park, Y., Kang, E., Kim, E. K., Kim, J. H., Kim, S. H., Suh, K. J., Kim, S. M., Jang, M., Yun, B. La, Park, S. Y., & Shin, H. C. (2022). Effect of Estrogen Receptor Expression Level and Hormonal Therapy on Prognosis of Early Breast Cancer. *Cancer Research and Treatment*, *54*(4), 1081–1090. https://doi.org/10.4143/crt.2021.890

Zeeberg, B. R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D. W., Reimers, M., Stephens, R. M., Bryant, D., Burt, S. K., Elnekave, E., Hari, D. M., Wynn, T. A., Cunningham-Rundles, C., Stewart, D. M., Nelson, D., & Weinstein, J. N. (2005). High-throughput GoMiner, an "industrial-strength" integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, *6*. https://doi.org/10.1186/1471-2105-6-168

Zhou, C., Huang, Y. Q., Da, M. X., Jin, W. L., & Zhou, F. H. (2023). Adipocyte-derived extracellular vesicles: bridging the communications between obesity and tumor microenvironment. In *Discover Oncology* (Vol. 14, Issue 1). Springer Science and Business Media B.V. https://doi.org/10.1007/s12672-023-00704-4

Zhou, T., Yao, J., & Liu, Z. (2017). *Gene Ontology, Enrichment Analysis, and Pathway Analysis*. http://www.geneontology.org/